Programming III Presentation Script

**Question 22**

The decision tree model was then evaluated using the Binary Classification Evaluator. The accuracy of these predictions was expressed as an area under the ROC curve. The decision tree model obtained an area under the ROC curve of 0.97.

**Question 23**

The same steps were conducted to predict fraud using the Random Forest Classifier. The classifier was imported from pyspark, the model was instantiated, the model was fit to the training data, predictions were made on the test data, and the area under the ROC curve was obtained.

**Question 24**

Finally, this process was repeated for the Gradient Boosted Trees algorithm.

**Question 25**

The AUC results for the 3 models were as follows:

Decision Tree: 0.97

Random Forest: 0.87

Gradient Boosted Trees: 0.99

**Conclusions and Recommendations**

Out of the 3 models run on this data, the Gradient Boosted Trees method had the highest accuracy, at 99%. This was followed by Decision Trees at 97%, and then Random Forest at 87%. We therefore recommend the organization proceed with implementing the Gradient Boosted Trees model to predict fraud.

It would be helpful to perform further analysis on all transactions coming from accounts that were found to have at least one fraudulent transaction. Data analysis on these accounts would identify overall trends in fraudulent activity. This information would further help to improve the fraud prediction models and to stop fraudulent activity more rapidly.

Future analysis could also focus on timestamps associated with transactions. It would be helpful to determine whether fraudulent transactions happen more at a given time of day or day of the week. Transactions happening during these time periods could be given extra scrutiny.